DOCUMENT RESUME

ED 212 165 FL 012 771

AUTHOR Woodford, Protase
TITLE A Common Metric for Language Proficiency. Final Report.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Department of Education, Washington, D.C.
PUB DATE Dec 81
GRANT G008001739
NOTE 39p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Testing; Higher Education; *Language Proficiency; *Language Tests; Listening Comprehension; Needs Assessment; *Norm Referenced Tests; Norms; Program Proposals; *Scaling; Secondary Education; *Second Language Learning; Speech Communication

ABSTRACT

This is a report on a project established to develop a "common yardstick" to describe performance in one or more language skills. Descriptive scales for oral interaction were prepared as well as a general outline of scale characteristics for listening comprehension and reading. Experts in the field reviewed the project proposal and recommendations were made to devote the major effort to oral interaction. The major outcomes at this stage were the following: (1) a commitment to some form of the 0-5 government scale; (2) concentration on the relationship between linguistic ability and the larger area of inter-personal communication; and (3) concentration of efforts at the 0-2 range, the one most second language speakers can expect to attain after the ordinary academic course of study. The major outcomes of the study are summarized as follows: (1) consensus on the usefulness of the expanded definitions at Levels 0 and 1; (2) agreement on the usefulness of a bilevel system; and (3) the need for definitions at Levels 0 and 1; (2) agreement on the coordination of efforts among the various agencies concerned with language proficiency testing. Immediate and long-range development work stemming from the "Common Yardstick" project is described by way of conclusion. (AMH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Title VI, NDEA

ED212165

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

FINAL REPORT

PROJECT TITLE: A COMMON METRIC FOR LANGUAGE PROFICIENCY

GRANT NO: G008001739

PROJECT DIRECTOR: PROTASE WOODFORD

SUBMITTED: DECEMBER, 1981

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541

FL012771

TABLE OF CONTENTS

Section		Page
1.	Background of Project	1
2.	Brief Description of Project Tasks	2
3.	Project Outcomes	3
	Task 1: Review by Measurement Specialists	3
	Task 2: Measurement Specialist Working Group Meeting	3
	Task 3: Final Scale Revision Meeting	7
4.	Summary and Further Planned Development Work	9

Appendices

A. Draft Scale Descriptions

B. Questionnaire on Draft Common Yardstick Scales

C. Final Level Descriptions for Oral Interaction

1. Background

As an outgrowth of the long-standing involvement of Educational Testing Service's language staff with the Foreign Service Institute (FSI) proficiency rating scale, and in response to expressions of interest from the British Council, the English Speaking Union, and German and Japanese agencies, ETS, in June 1979, sponsored a small conference to discuss the possibility and desirability of establishing a "common yardstick" (or yardsticks) to describe performance in one or more language skills.

At the conference, descriptive scales of language proficiency developed in various countries and by international agencies such as the Council of Europe were distributed. Both theoretical and practical issues in the development and use of a single set of descriptive scales on an international basis were discussed. Background papers developed from prior work in this area were presented by ETS and by British Council staff, and other participants contributed information from the perspective of their own organizations.

There was unanimous agreement among the participants that development of descriptive scales for all language skill areas should be attempted. It was also recommended that a small working group from among the conference participants be designated to begin work on the scale development.

In November, 1979, a working group consisting of John Clark and Protase Woodford of ETS; Brendan J. Carroll, British Council; David Hicks, English Speaking Union; and Anthony Fitzpatrick, Deutscher Volkshockschul Verband met in London. The outcome of this meeting was the preparation, in rough draft form, of descriptive scales for oral interaction and writing, and a general outlining of scale characteristics for listening comprehension and

reading. Following the November meeting, draft scale descriptions for "passive" listening comprehension (excluding oral interaction) and for reading comprehension were prepared. These draft scales are shown in Appendix A.

On the basis of these initial activities and the positive general response obtained, ETS requested and received funding from the foreign language and area studies research program (U.S.O.E./D.E.) for the current project to continue work on a common metric for language proficiency.

2. Brief Description of Project Tasks

Proposed further activities for this project included the following three tasks:

(1) Distribution of the draft scales to recognized foreign language measurement specialists for their critique, commentary, and any suggestions for revision.

(2) Convening of a small group of measurement specialists to synthesize the recommendations of the reviewers and collaborate in the revision of the scales. The senior British member of the international working group would also be invited to attend this meeting to provide a summary of similar inputs by measurement specialists in Europe.

(3) Assuming a generally positive outcome for activities (1) and (2), presentation and discussion of the language assessment scales and recommendations for future development activities to implement the use of these scales to executive officers of foreign language associations, government agency representatives, and representatives of the international business community.

3. Project Outcomes

Task (1): Review by Measurement Specialists

In December, 1980, selected foreign language measurement experts were sent project information and draft scales for their review and comment. Individuals requested to participate in this review included:

Dr. Lyle Bachman
University of Illinois

Dr. Michael Canale
Ontario Institute for Studies in Education

Dr. James Child
National Security Agency

Dr. Ray Clifford
Central Intelligence Agency and
Defense Language Institute

Dr. James R. Frith
Foreign Service Institute

Dr. Barbara Freed
University of Pennsylvania

Dr. Helen Jorstad
University of Minnesota

Dr. Pardee Lowe
Central Intelligence Agency

Dr. Adrian Palmer
University of Utah

Dr. Howard Nostrand
University of Washington

Dr. G. Richard Tucker
Center for Applied Linguistics

Reviewers were asked to give (1) their appraisal of the overall merit of the project from both psychometric and practical standpoints, and (2) specific suggestions for the revision of the draft scales with the rationale for such revision. The questionnaire prepared by project staff to collect this information is included as Appendix B of this report.

Task (2): Measurement Specialist Working Group Meeting

From the original group of reviewers listed above, a smaller working group was selected to participate in an intensive two-day meeting at ETS to consider the comments of all reviewers and to collaborate with ETS

staff on the revision of each of the four language scales. The following individuals participated in the February 24-25, 1981 meeting at ETS:

Protase Woodford, Project Director
John Clark, Principal Investigator
Judith Liskin-Gasparro, Associate Examiner
Marianne Adams, Foreign Service Institute
Lyle Bachman, University of Illinois
Michael Canale, Ontario Institute for Studies in Education
James Child, National Security Agency
Ray Clifford, Defense Language Institute
Barbara Freed, University of Pennsylvania
Pardee Lowe, Central Intelligence Agency
Howard Nostrand, University of Washington

At the February 24, 25 meeting, the participants discussed in detail the following issues:

1. Skills Represented by Scales

It was suggested that the traditional four skills--listening, speaking, reading and writing--be modified to include reading, writing and "pure" listening comprehension as discrete, measurable skills, and that "oral interaction" be used to replace "speaking" because of the listening skill required in real-life speech contexts.

2. The number of Scale Divisions

The scales reviewed in connection with the project came from a variety of U.S., British, and European sources. The 0-5 Foreign Service scale was the most familiar to the participants. The utility of the various scales was discussed as well as a proposed 8-level scale (0-7) for oral interaction.

3. FSI Scale/oral interaction scale comparisons

The relationship of the proposed 0-7 oral interaction scale to the FSI 0-5 (with +'s) scale was considered.

4. Scale Progression

Participants were asked to consider whether the proposed scales provided for a smooth progression from one level to the next and whether any pair of level descriptions was too close to allow for a meaningful distinction between levels.

5. Intra-level consistency

Participants were asked to consider whether the descriptive statements within each level would apply to most persons within the ability group; i.e. to make sure that there would be no descriptions of tasks or behaviors "too easy" or "too difficult" for people within the level.

6. Inter-scale comparability

Discussion centered on the degree to which the scales for the four skills were consistent with regard to detail of description.

7. Individual Scale Aspects

Participants were asked to rank each of the four draft scales on the following criteria:

A. "understandability"

B. "real-life referencing"

C. ease and straightforwardness of use for rating examinee performance

D. priority for development.

The participants were also asked to consider the scales presented to them in light of the "ideal" scale; that is, one that would include all of the features of oral interaction that they considered important.

As the meeting developed, it became apparent that the task at hand was extremely complex. Consequently, it was decided that a major part of

the effort would be devoted to oral interaction.

The major outcomes of the meeting were the following:

(1) A commitment to some form of the 0-5 government scale. The deliberations of the group demonstrated that all of the members were in some sense basing their reactions to the draft scales on the relationships of these newer scales to the government scale developed by the FSI. Since the government scale is relatively better known and since it has a long and respected history, it seemed most reasonable to begin with it as a base, making adjustments to it that would not alter the accepted understandings of the significance of Level 1 proficiency, Level 2 proficiency, etc.

(2) The realization that no scale currently in existence or under consideration does as complete a job of evaluating oral proficiency as the participants in the February meeting would like. Particular concern was focused on such aspects of language ability as register, cultural sensitivity, and in general the relationship between linguistic ability and the larger area of inter-personal communication. While these issues arise mostly at the upper proficiency levels, there are some languages for which they emerge as low on the government scale as Level 2. Time was also spent discussing and coming to a common understanding of the term "fluency."

(3) The decision that further work is most essential at the 0-2 range. This is the area in which most second-language speakers can expect to fall after taking advantage of the range of academic courses and extra-curricular activities usually offered in secondary schools and colleges. Level 3 proficiency is usually attained only after extended residence in a country where the target language is spoken and/or through intensive

or immersion-type language study. It was recommended that Levels 0, 1, and perhaps 2 be further subdivided to provide finer distinctions.

Given the outcomes of the February meeting, it became apparent that further refinement of the scale descriptions was needed before proceeding to the expected next step of the project, the convening of a meeting with executive officers of foreign language associations, government agency representatives, and representatives of the international business community. The intended focus of a meeting with these "user groups" was planned to be a presentation of the revised scales and a discussion of about whether and to what extent the scales met their specific needs in the area of language proficiency evaluation.

Since the group at the February 24-25 meeting had recommended that further work be done on the lower levels of the oral interaction scale and, further, that work on the other skills be postponed while efforts were focused on the oral interaction scale, it was decided not to hold the meeting for language association officers, government agency representatives, and representatives of the international business community as planned. Instead, further work was done at ETS on the expansion of the lower end of the oral interaction scale (see Appendix C).

Task (3): Final Scale Revisions

On October 6, 1981, the final meeting of the Common Yardstick Project was held at the CIA Language School, hosted by Dr. Pardee Lowe. The purpose of the meeting was to discuss and, if possible, reach consensus on the revisions to the oral interaction scale and to agree on future plans. The participants at the meeting were as follows: Protase E. Woodford, Judith E. Liskin-Gasparro, and Ihor Vynnytsky from ETS; Professor

Barbara Freed, Assistant Dean for Languages, University of Pennsylvania; Dr. Ray Clifford, Academic Dean, Defense Language Institute; Dr. Pardee Lowe, CIA Language School; Dr. Yvonne Escolá, teacher of French in Montgomery County, Maryland and program officer, National Endowment for the Humanities, and Dr. John L. D. Clark, Center for Applied Linguistics.

The discussion at the meeting focused on the needs of the government language schools and the academic community in the area of the evaluation of oral proficiency. The expanded lower end of the government scale was presented for discussion, and both Dr. Clifford and Dr. Escolá agreed that it would provide valuable information for students as well as teachers of language programs. At the CIA Language School, language testers often offer finer descriptive distinctions, beyond the official ratings, in their evaluations of student examinees. Dr. Lowe reported that the informal descriptions were very similar to the expanded "intra-level" descriptions prepared by ETS. This congruence in the independently developed descriptions was an encouraging sign, and it was agreed that further work in this area would benefit from the experience of the CIA Language School.

The group discovered a second area of congruence between academic and government language assessment needs in the discussion of the value of a bi-level system of oral proficiency assessment. It was agreed that for other than a few very specialized uses, there is little need to discriminate between levels of proficiency at the 3+–5 range. Most academic and professional needs will be satisfied by proficiency at the 3 level or lower, so it is at this lower end of the scale that most attention needs to be focused. The group agreed that the following labels for ranges of proficiency represent realistic descriptions:

0 elementary

1 intermediate

2 advanced

3-5 superior

In summary, the major outcomes of the meeting were as follows:

(1) Consensus on the usefulness of the expanded definitions at Levels 0 and 1. It was agreed that these descriptions were definitely "on the right track," and that further work on them would be a logical, and valuable, next step.

(2) Agreement on the usefulness of a bilevel system, according to which further development efforts would be concentrated in the 0-3 range. Individuals above Level 3 would be designated as "superior." If a precise level were desired, it could be provided via the traditional face-to-face interview.

(3) Agreement that coordination of efforts among the various agencies concerned with language proficiency testing is a major concern that must be addressed. As of this date, a major cooperative venture, stemming from the Common Yardstick Project, has been launched that includes ACTFL; ETS; and the CIA Language School. (See page 10 below for further discussion.) In addition, an invitational conference on language proficiency testing was hosted by Ray Clifford and the Defense Language Institute November 30-December 1, 1981 in order to discuss the current needs and projects of agencies inside and outside the government, and to decide on areas of future development.

4. Summary and Further Planned Development Work

Although the final stage of the project, i.e. the meeting with representatives of language associations, government agencies, and internal

business, did not take place as proposed, the Project Director feels that the new direction undertaken by the project will serve to build a stronger foundation to serve these constituencies better in the long run. The contributions of the language professionals and linguists were valuable in defining areas of strength and weakness in the existing scales, and especially in recommending that new efforts be based on the government definitions and scale. The decision to concentrate on the lower end of the oral interaction scale resulted in the creation of intermediate working definitions. For several years language professionals in academe have recognized that the absence of these expanded descriptions has severely limited the applicability of the government and interaction scale to college and high school students.

Further development work, stemming from the Common Yardstick project, is already underway by ACTFL, ETS, and the CIA Language School. After the October 6 meeting, Mr. Woodford turned over to ACTFL the expanded descriptions of oral proficiency for Levels 0 and 1. ACTFL, which is working on the development of proficiency levels as goals of instruction under a grant from the International Research and Studies of the U.S. Department of Education, in turn asked Dr. Lowe of the CIA Language School to investigate the validity and accuracy of the ETS descriptions. Dr. Lowe, assisted by funds and professional collaboration from ETS, designed a research project to determine (1) whether the expanded intra-level descriptions correspond to real-life language use; and (2) whether the intra-level descriptions and independent raters will rank a group of tapes known to be within a given level in the same order. The scale with the expanded lower end will be taught to college faculty members in Spanish and French at the workshop sponsored by ACTFL and conducted by ETS under a grant to ACTFL by the U.S.

Department of Education.

For the long range, development work similar to that which has been accomplished for oral interaction might be undertaken for the other language skills. Although further development work beyond the scale-definition and review stage would require additional financial support, and would also, of course, depend on the psychometric appropriateness and anticipated practical utility of the final scale descriptors, a fairly large-scale test development/validation project could be envisioned as a possible outcome of the initial work. This larger study could include each of the following activities: (1) Development of comprehensive measures of language skills, encompassing and operationally defining the descriptive scales. These would be very exhaustive and lengthy direct measures of each of the skills in question, requiring perhaps two full days of testing on the part of each examinee. It is recognized that these criterion tests would not be practical for regular measurement purposes, but would be used as comprehensive "benchmark" instruments exemplifying the scale descriptions and against which presently available, more easily administered tests (or smaller-scope tests yet to be developed) could be compared and validated.

(2) Development of validation measures external to both the large-scale "benchmark" tests and any smaller-scope tests. These external measures would be expected to include both examinee self-appraisal and "second-party" (e.g., classroom teacher, work supervisor) evaluation of the examinee's proficiency in the language skill areas at issue. These evaluations could take the form both of (a) direct utilization of the common yardstick scales (i.e., examinees and second-party observers would be asked to rate the performance vis-a-vis the common yardstick

descriptors); and (b) use of more detailed and more "atomistic" descriptions of particular language-use functions (e.g., "say the days of the week," "buy clothes in a department store," "talk about my favorite hobby at some length, using appropriate vocabulary"), which would be rated on a dichotomous (can do/cannot do) basis.

(3) Large-scale administration of the comprehensive "benchmark" measures, smaller-scope measures, and external criterion measures to a large and varied group of examinees, for purposes of both construct/concurrent validation of the instruments in question and establishment of equating data relating examinee performance on the smaller-scope tests to both the "benchmark" test results and to the common yardstick descriptors.

The exact nature and operational details of the activities outlined in 1-3 above would, of course, have to be spelled out much more comprehensively at a later date; the intent at this point is simply to give a general overview of the kinds of development work that would seem to be logical and, we hope, practically useful extensions of the initial development of the common yardstick descriptors.

Conclusion

At the time that the current study was proposed, the idea of a "common yardstick" or uniform descriptors of language proficiency was being considered only within a restricted population of measurement specialists and government connected linguists. The "yardstick" activities themselves and the reports on the yardstick to major foreign language education constituencies* have created extraordinary interest in

---

*Clark, Freed, Liskin-Gasparro, Lowe, Woodford have reported on the "Yardstick" to such groups as: Southern Conference on Language Teaching, Modern Language Association of America, Pennsylvania MLA, Florida Foreign Language Teachers' Association.

the project across all academic levels. It is obvious now that the original scope of this project was far too broad. A result of the deliberations at the February meeting of the working group was a narrowed focus on a scale for one of the skill areas, the one considered of highest priority--oral interaction. The well-known and respected scale used by the federal government has seen limited use in the academic context primarily because it provides too little discrimination at the lower end 0.0-2.0. It is precisely at the lower end of the scale where there is greatest need for evaluation of language skills in schools and colleges.

The proposed expanded scale is a product of the working group's efforts subsequent to the February 1980 meeting and during the October 1980 meeting in Arlington, VA to refine and plan next steps.

The work accomplished through this project has served and will continue to serve a number of related projects.

The working group members are actively involved in continued dissemination of the draft scale to various foreign language constituencies. The expanded oral interaction scale is currently undergoing validation under an ACTFL-sponsored project and will--if proved valid--likely become the "Yardstick" for describing the ability of American students to function in a real-life communication situation. Further work on the "Yardstick" including further development and refinement of the existing draft scales for "pure" listening comprehension, reading and writing is planned. Support will be sought from a variety of sources. Training programs for high school and college foreign language teachers are scheduled for 1982. These training programs will focus on the evaluation of students' ability

to understand and speak in a foreign language in a real-life context. The scale considered for use is the expanded, revised oral-interaction scale developed under the current Common Yardstick project. Among the sponsors of the training programs are:

The American Council on the Teaching of Foreign Languages (ACTFL)
The Northeast Conference on the Teaching of Foreign Languages
Educational Testing Service
Vassar College

APPENDICES

A. DRAFT SCALE DESCRIPTIONS

B. QUESTIONNAIRE ON DRAFT COMMON YARDSTICK SCALES

C. LEVEL DESCRIPTIONS FOR ORAL INTERACTION

Level Descriptions for Oral Interaction

LEVEL	
0	No functional communication in the language.
1	Speech limited to short utterances, in large part formulas (very limited repertory). Requires cooperative, sympathetic interlocutor. Can deal only with highly predictable transactional situations and is baffled by non-predictable elements. Extensive use of paralinguistic communication is required. Except for memorized expressions struggles to make every utterance. Pronunciation is generally intelligible though clearly non-standard. Virtually no grammatical control except in stock phrases.
2	Can express some simple information and ideas in addition to stock formulas, but speech is still limited to short utterances. Relies heavily on well-rehearsed sentence patterns and requires cooperative, sympathetic interlocutor. Utterances still made with great effort. Pronunciation is clearly non-standard but generally intelligible. Except in stock phrases, there are repeated errors in basic constructions.
3	Is able to dicuss situations, relevant to his own situation, in the form of simple dialogues. Utterances can be longer and more connected than in level 2. Can respond in a limited manner to non-routine questions. Unlikely to initiate new conversational topics. Speech is usually hesitant. Frequent grammatical errors, but fair control of basic constructions.
4	Can handle most communication relevant to his situation in a spontaneous manner. Can initiate new topics of discussion and express opinions in a simple manner. Speech is occasionally hesitant. Makes few errors in basic constructions but has some difficulties with more complex syntax.
5	Can maintain conversation on most formal and informal topics, including communication of some abstract concepts. Shows good independence in discussion, needing only occasional assistance of interlocutor. Can react appropriately to rapid change of topic. Is occasionally hesitant. Makes occasional grammatical errors, but has generally good control of grammar. Vocabulary is extensive.
6	Can express himself appropriately on a wide variety of topics with fluency, precision, and appropriate register. Can maintain his own part in conversation effectively. Pronunciation is occasionally slightly non-standard. Vocabulary is very extensive. Makes only very infrequent grammatical errors of a type that would not be expected of an educated native speaker.
7	Can express himself entirely appropriately and effectively with grammatical accuracy and easy fluency.

Level Descriptions for Listening Comprehension

LEVEL	
0	May catch an occasional spoken word or formulaic expression, but no functional understanding of stream of speech.
1	Partial comprehension of clear train departure, other similar announcements. Can understand a few key words of "special English" broadcasts and "tour guide" speech.
2	Good comprehension of clear train departure, etc. announcements. Partial comprehension of special English, tour guide, and other situations involving careful and somewhat deliberate speech. Partial comprehension of slow, carefully enunciated, and simplified telephone speech.
3	Virtually complete comprehension of clear loudspeaker announcements, special English broadcasts, and "tour guide" situations. Can get the gist of factually oriented news broadcasts. Has some idea of the content of lectures and other formal presentations in subject areas with which he or she is familiar. Has reasonably complete comprehension of slow and careful telephone speech. Can detect major affective components of speech (e.g., anger, incredulity).
4	Essentially complete comprehension of factual news broadcasts. Partial comprehension of news commentary and analysis. Reasonably good comprehension of formal presentations in familiar subject areas. Can partially understand movie sound tracks, stage plays, other dramatic arts presentations. Reasonably good comprehension of telephone conversation using educated speech at normal delivery rates. Can catch some of the words of popular songs. Can get the gist of an overheard conversation between native speakers, provided they are not speaking rapidly or colloquially. Can understand non-native speakers of the language when they are speaking slowly and carefully. Can get the gist of regionally accented speech and/or speech using regional vocabulary expressions. Can understand non-colloquial children's speech. Can usually detect emotional tone of speech, including irony, sarcasm, etc. as well as the more basic affective elements.

5 Reasonably complete comprehension of news commentary and analysis, lectures on a variety of topics outside of area of specialization.
Reasonably good comprehension of movie sound tracks, plays, and other dramatic presentations, provided that actors are not speaking rapidly or colloquially.
Good telephone comprehension of normally speeded, educated speech.
Can get the general theme of most popular songs with careful listening.
Reasonably good comprehension of overheard conversations between educated speakers.
Reasonably good comprehension of non-native speakers, regionally accented speech, and children's speech.

---

6 Virtually complete comprehension of news commentary and analysis, lectures, and other formal presentations.
Very good comprehension of movies, plays, dramatic arts presentations.
Virtually complete comprehension of all telephone conversations, except for highly colloquial or extremely speeded.
Good comprehension of non-native, regional, and children's speech.
Virtually complete comprehension of overheard conversations.
Can isolate and generally understand a particular overheard conversation in "cocktail party" situations.

---

7 Listening comprehension closely approaching the performance of educated native speaker. Only observable deficiencies are occasional non-comprehension of highly idiomatic expressions, extremely regional speech, or conversations between native speakers deliberately intending to conceal the content of their conversation. Has full functional comprehension of all speech situations normally encountered in everyday and professional contexts.

Level Descriptions for Reading Comprehension

LEVEL	
0	Can discriminate some of the orthographic characters in the language and make out an occasional word; essentially no functional comprehension of the printed language.
1	Can comprehend simple street signs, store front designations ("restaurant," "post office," etc.). Essentially no comprehension of sentence-length printed material, even of an intentionally simplified nature.
2	Can make out the general topic of extremely simple texts, but lexical and grammatical deficiencies give rise to some confusion. Constant dictionary or glossing support required.
3	Can read fairly fluently, and with only occasional recourse to a dictionary, personal letters or notes in which the writer has deliberately used simple lexicon and structure. Can get the general drift of relatively straightforward newspaper or newsmagazine prose, but cannot usually follow the factual message on a sentence-by-sentence basis (i.e., without frequent blocks to comprehension). Is insensitive to tone and other affective aspects of the text. Is not sensitive to passage tone, and attends only to informational message.
4	Can read with reasonable fluency personal letters or notes written as they would be to a native user of the language. With occasional use of dictionary, can understand typical newspaper/newsmagazine articles on a sentence-by-sentence basis, except for articles in highly technical areas (e.g., banking and finance, political commentaries). Attempts reading of novels or other lengthy narrative material but frequency of recourse to dictionary makes this task very tedious.
5	Can read most texts on non-technical subjects with good comprehension and only occasional use of dictionary; exhibits similar level of performance in technical areas with which he/she is familiar. Where present in text, can detect major elements of affect (anger, disbelief, etc.) on part of author. Frequently fails to understand or misinterprets cultural references, highly colloquial usages.

6 Reading speed and level of comprehension at a high level for all material normally encountered in popular press and/or areas of his/her specialization, with no or very infrequent use of dictionary. Only occasional misunderstanding of cultural references or colloquial expressions, and can appreciate subtle affective aspects of the author's expression (irony, sarcasm, etc.).

---

7 Reading performance virtually indistinguishable from that of educated native speaker, with only highly infrequent and subtle indications of non-native performance in areas such as comprehension of telegraphic writing (e.g., newspaper want ads) and of generally-known but rarely employed cultural references. For all practical language-reception purposes, can be considered functioning at a "native speaker" level.

Level Descriptions for Writing

LEVEL	
0	Either cannot write any of the language at all or, at most, only a few letters or isolated words. No communication.
1	Able to string words together. Little grammatical accuracy. Impoverished lexis. Message is comprehensible only in parts and is confused and unclear in others.
2	Has just sufficient control of vocabulary and grammar to express the bare essentials of the message simply. There will be many inaccuracies of usage and many inappropriacies of style. Lacking in range of skills, flexibility of presentation and the use of suitable cohesive features. Conveys the essential message but not much more; however, this is the first real stage in communciation in writing.
3	Despite some vocabulary and grammatical inaccuracies and inappropriate usage, can use basic structures and vocabulary to convey his meaning. Cohesive features, where attempted, contribute little to the coherence of the discourse. The basic message is conveyed, although certain details are not clear.
4	Usually accurate in grammar and adequate in use of vocabulary. There are occasional errors but they do not create major problems in understanding. Lacking in flexibility of style. Range of skills appropriate to the purpose. Cohesive features are attempted but do not always assist the structure of the presentation. Content of the message is conveyed clearly but the attitude of the writer is not always discernible.
5	Control of grammar is adequate for the purpose with only occasional inaccuracies. Ample range of appropriate vocabulary but little flexibility of style; some use of appropriate cohesive features. Writer conveys the message with clarity but may not be able to indicate his attitude or other subleties.
6	Accurate use of grammar with a wide range of vocabulary and proper use of cohesive features. Somewhat lacking in flexibility for given writing tasks. Style and range of skills basically appropriate to the purpose of the message. The purpose and content of the message are clear although some deeper aspects of other's opinions may not always be so clear.
7	Completely accurate use of grammar and vocabulary. Register, style and method of presentation entirely appropriate. All use of cohesive features. All aspects of the message are conveyed very effectively.

Questionnaire on Draft Common Yardstick Scales

BACKGROUND

As described in the accompanying letter, the four descriptive scales are an initial and at the moment very rough attempt to designate levels of general proficiency that would have close congruence with observed language performance by second-language learners at various stages of language acquisition and would at the same time be simple, straightforward, and easily interpretable by non-specialist users of the scale information.

Development of suitable scale descriptions is really a circular (we hope not viciously circular) process. On the one hand, it would appear necessary to develop at least very tentative initial descriptions of sequential performance levels in order to have some conceptual basis to work from in developing assessment procedures to measure these performances. On the other hand, unless and until measurement techniques are developed that are capable of measuring in a very straightforward way the performances associated with a given level, it is not possible to validate these descriptions as reflective of empirically observed language performance (or to modify them as necessary to produce the desired congruence with the "real-world" data obtained).

As the initial step in this process, we have considered it desirable to begin with tentative (and at present hypothetical) descriptions of language performance, and to request the assistance of others who have been closely involved with other relevant measurement studies and projects to modify and refine the draft scales in light of their own experiences in these areas. If we can prepare tentative scales that are considered to have useful psychological/psychometric/practical meaning as judged by the general experience of persons who have been working closely in these areas, it will then be possible to develop measurement instruments exemplifying the draft descriptions and administer these on a fairly large-scale basis as guides to the modification and validation of the descriptions.

For each of the questions below, we would therefore request your own best judgment, on the basis of your prior familiarity with language proficiency rating scales and related testing activities, with the understanding that the preparation of <u>draft scale descriptions</u> is only the first step in the entire scale development and validation process. Space is provided in the questionnaire for additional discussion of individual questions, as well as for further comments and suggestions about the common yardstick project as a whole and the test development/validation activities that would be expected to follow the initial draft scale specifications.

SKILLS REPRESENTED BY SCALES

As you know, the "traditional" division of language skills into the four categories of listening comprehension, speaking, reading, and writing often fails to coincide with actual language-use situations, in which more than one of these skills are exercised simultaneously. In these situations, it becomes

both difficult and potentially misleading to attempt to measure the different skill components as separate entities.

ORAL INTERACTION

In view of these considerations, we have adopted the term "oral interaction" to represent the combined production/reception skills involved in live, face-to-face communication with a speaker of the target language, and the individual level descriptions for "oral interaction" include both production and reception aspects.

To reflect and measure the listening comprehension skills associated with "one-way" communication, in which the examinee listens passively to radio or television broadcasts, loudspeaker announcements, stage plays, overheard speech, lectures, etc., the "listening comprehension" scale is based only on these types of aural comprehension situations. From a language-learning standpoint, a division between "oral interaction" and (one-way) "listening comprehension" ability is considered useful, since it allows for situations in which the language-training emphasis is on "message reception" (e.g., monitoring radio broadcasts; listening comprehension for personal enjoyment of foreign language programs) rather than face-to-face communication in the language.

Although there are situations in which reading comprehension and writing activities take place more or less simultaneously (as in taking written notes on a textbook chapter), the interaction is by no means as frequent or pervasive as is the case in listening/speaking. It was thus considered appropriate and desirable to address these two skill areas separately as "reading comprehension" and "writing proficiency."

Do you feel that developing performance descriptions in the four described areas of oral interaction, (one-way) listening comprehension, reading comprehension, and writing proficiency is an appropriate way to "divide up" the language skills for assessment purposes, without doing violence to the way these skills are called upon in real-life language use? Any additional suggestions or other comments in this regard?

NUMBER OF SCALE DIVISIONS

With regard to the number of divisions (levels) on the descriptive scales, it was intended to provide enough levels to permit a usefully wide range of descriptions while at the same time not being so "fine-grained" that (a) raters would not be able to reliably distinguish between adjacent scales and/or (b) the difference in language performance represented by adjacent scales would not reflect a pragmatically useful difference in examinee performance.

The FSI oral interview scale (also enclosed for comparison purposes) consists of 11 points if "plus" values are included, or 6 points when only the whole levels (0, 1, 2, 3, 4, 5) are considered. We have experienced some scoring variability in the 11-point FSI scale, and generally much more consistent rating based on the whole-level scale (i.e., scoring only within whole levels and not attempting to assign plus values).

The proposed scale for oral interaction consists of 8 points (0-7), which is intermediate between the FSI whole-level and "with-plus" scales. It should also be noted that the highest level (7) on the proposed scale does not represent educated native speaker proficiency but highly proficient non-native (high 4 or 4+ on FSI scale). We would appreciate any observations on the proposed number of scale divisions for the oral interaction scale, as well as for the other proficiency scales.

FSI SCALE/ORAL INTERACTION SCALE CROSS-COMPARISONS

Since the FSI interview scale is so well known and so extensively used, with very satisfactory results, in a number of different proficiency measurement applications, the official FSI level description definitions were very carefully considered in drafting the oral interaction scale descriptions. We would appreciate any comments or comparisons you could make concerning the level descriptions of the draft oral interaction scale vis-a-vis the FSI descriptions (e.g., degree of specificity, ease of interpretation) or other aspects of the two scales.

SCALE PROGRESSION

Based on your knowledge of typical language learning sequences, does each of the four draft scales appear to show a logical and smooth progression from one level to the next, or are there discontinuities (too quick a performance jump between two given levels)? Alternatively, are the descriptions for any pair of adjacent levels so close that they fail to describe any real differences in performance? (Please answer this question in terms of each proficiency scale considered separately.)

INTRA-LEVEL CONSISTENCY

Although there will of course be variations in the exact profile of examinee language performance, depending on language learning history, it is hoped that the descriptive scales will be generally applicable to most language learners, in the sense that the individual statements comprising a given level description will all be applicable to most persons in that general proficiency grouping. Alternatively stated, a given level description should contain no individual elements that are either "too easy" (insufficiently stringent) or "too advanced" (too demanding) for a given level, by comparison to the other components of the description given for that level. Please identify any individual elements within any of the level descriptions for each of the four scales that seem "out of place" for that level.

INTER-SCALE COMPARABILITY

It is not intended to "equate" the four proficiency scales in the sense of being able to say, for example, that level 3 for oral interaction corresponds to level 3 listening comprehension for typical language learning curricula (although it would of course be possible to eventually obtain relevant information in this regard). However, in order to provide reasonable uniformity in the type and amount of descriptive detail included in each of the scales, some attention to across-scales consistency is needed. We would appreciate your observations on the relative degree of detail provided in each of the scales (on an across-scales comparison basis), as well as your suggestions on whether a given scale or scales should be fleshed out in greater detail, made more abbreviated or more general, or otherwise revised, in general comparison to the set of scales considered as a whole.

INDIVIDUAL SCALE ASPECTS

We would appreciate your rank-ordering (1=best) of each of the four proficiency scales, in their present draft form, on each of the following dimensions. Please make a choice, even if "forced," in each instance.

(A) "Understandability" of the scale descriptions for lay users of the scoring information (e.g., employers, admissions officers not in the language field, parents of language students, etc.):

_____ Oral Interaction

_____ Listening Comprehension

_____ Reading Comprehension

_____ Writing Proficiency

Additional Comments?

(B) Degree of "real-life referencing" of the scale descriptions (i.e., extent to which the descriptions refer to actual language-use tasks, as opposed to linguistically-based criteria):

_____ Oral Interaction

_____ Listening Comprehension

_____ Reading Comprehension

_____ Writing Proficiency

Additional comments on the preceding?

(C) Ease and straightforwardness with which the scale could be used in rating examinee performance (assuming appropriate prior training of the raters):

_____ Oral Interaction

_____ Listening Comprehension

_____ Reading Comprehension

_____ Writing Proficiency

Additional comments relating to (C)?

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

(D) If it were not possible, for whatever reason, to fully develop each of the descriptive scales simultaneously (including the related test development and validation activities), what priority ranking for development would you give each of the four scales?

_____ Oral Interaction

_____ Listening Comprehension

_____ Reading Comprehension

_____ Writing Proficiency

OTHER

With respect to the general scale development/instrument development/ administration/analysis/revision process described in the cover letter (and more fully discussed, with respect to speaking testing, in the "Toward a Common Measure..." paper enclosed), does this appear to be a useful and psychometrically appropriate overall procedure for defining and validating meaningful levels of language proficiency and the associated assessment instruments? Both general comments on the research strategy and more specific procedural suggestions would be appreciated.

Without commitment at this juncture, would you be interested in helping us to pursue the scale development and validation project further, through additional correspondence and/or working meetings, as appropriate?

_____

As the final item, the space below is for any further discussion of any aspects of the project plan and/or the draft scale descriptions that ar not sufficiently well addressed in previous questions. We very much appreciate your assistance in this effort, and will plan to keep you closely in touch on further project activities and results.

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

LEVEL DESCRIPTIONS FOR ORAL INTERACTION

Level 0: No ability to understand or speak the language.

Level 0A: Unable to function in the spoken language. Oral production is limited to occasional isolated words. Essentially no communicative ability.

Level 0B: Able to operate only in a very limited capacity within very predictable areas of need. Vocabulary limited to that necessary to express simple elementary needs and basic courtesy formulae. Syntax is fragmented, inflections and word endings frequently omitted, confused or distorted and the majority of utterances consist of isolated words or short formulae. Utterances rarely consist of more than two or three words and are marked by frequent long pauses and repetition of an interlocutor's words. Pronunciation is frequently unintelligible and is strongly influenced by first language. Can be understood only with difficulty, even by persons such as teachers who are used to speaking with non-native speakers or in interactions where the context strongly supports the utterance.

Level 0C: Able to satisfy immediate needs using learned utterances. There is no real autonomy of expression, although there may be some emerging signs of spontaneity and flexibility. There is a slight increase in utterance length but frequent long pauses and repetition of interlocutor's words still occur. Can ask questions or make statements with reasonable accuracy only where this involves short memorized utterances or formulae. Most utterances are telegraphic and word endings (both inflectional and non-inflectional) are often omitted, confused or distorted. Vocabulary is limited to areas of immediate survival needs. Can differentiate most phonemes when produced in isolation but when they are combined in words or groups of words, errors are frequent and, even with repetition, may severely inhibit communication even with persons used to dealing with such learners. Little development in stress and intonation is evident.

Level 1A: Able to satisfy basic survival needs and minimum courtesy requirements. In areas of immediate need or on very familiar topics, can ask and answer simple questions, initiate and respond to simple statements, and maintain very simple face-to-face conversations. Almost every utterance contains fractured syntax and other grammatical errors. Vocabulary inadequate to express anything but the most elementary needs. Strong interference from L_1 occurs in articulation, stress and intonation. Misunderstandings frequently arise from limited vocabulary and grammar and erroneous phonology but, with repetition, can generally be understood by native speakers in regular contact with foreigners attempting to speak their language. Little precision in information conveyed owing to tentative state of grammatical development and little or no use of modifiers.

Level 1B: Some evidence of grammatical accuracy in basic constructions, e.g. subject-verb agreement, noun-adjective agreement, some notion of inflection. Vocabulary permits discussion of topics beyond basic survival needs, e.g. personal history, leisure time activities.

Level 1C: Able to satisfy all survival needs and limited social demands. Developing flexibility in a range of circumstances beyond immediate survival needs. Shows some spontaneity in language production but fluency is very uneven. Can initiate and sustain a general conversation but has little understanding of the social conventions of conversation. Limited vocabulary range necessitates much hesitation and circumlocution. The commoner tense forms occur but errors are frequent in formation and selection. Can use most question forms. While basic word order is established, errors still occur in more complex patterns. Cannot sustain coherent structures in longer utterances or unfamiliar situations. Ability to describe and give precise information is limited. Aware of basic cohesive features (e.g., pronouns, verb inflections), but many are unreliable, especially if less immediate in reference. Extended discourse is largely a series of short, discrete utterances. Articulation is reasonably comprehensible to native speakers used to dealing with foreigners, can combine most phonemes with reasonable comprehensibility, but still has difficulty in producing certain sounds, in certain positions, or in certain combinations, and speech will usually be labored. Still has to repeat utterances frequently to be understood by the general public.

SOME SUGGESTED ADAPTATIONS OF THE FSI SCALE WITH EQUIVALENT VALUES

5.0
4.5
4.0

"SUPERIOR" This rating connotes no pattern of error in speech with fully developed vocabulary and comprehension.

3.5 — — cut @ 3.0 for wider use

3.0
2.5
2.0

As currently described in the FSI Scale

1.5 Suggested 1-C

1.0 Suggested 1-B, 1-A

+0 0.5 Suggested 0-C, 0-B

0.0 Suggested 0-A, 0-0